Enhancing Histopathological Breast Cancer Image Classification using Deep Learning

Puspanjali Mohapatra, Baldev Panda, Samikshya Swain

***Abstract:** In this paper we have conducted experimental analysis in predicting IDC (Invasive Ductal Carcinoma) as well as well as Malignant and Benign tumors from textual and histopathology image datasets. The analysis commences with the conventional machine learning algorithms on the text dataset and upgrades to deep learning while dealing with histopathology images. The machine learning algorithms like Logistic Regression, SVM, KNN, Decision tree are applied on the datasets to compare the accuracy among them. The model giving the best accuracy is decided through Feature extraction techniques like PCA and LDA leading to an improvement in accuracy. When dealing with large datasets consisting of high-resolution images, the machine learning algorithms don't perform well. Deep learning has the ability to handle such complex situations which include high-dimensional matrix multiplications. Various architectures of CNN were applied and the model with the high generalization accuracy and minimal complexity is selected. The histopathology images are given as input to the CNN network as training models and then finally classified as having IDC or Malignancy. The best model is selected after varying the number of hidden layers and then applied to the dataset for final classification.*

***Index Terms:**Breast cancer, IDC, Histopathology Images, Machine Learning, Logistic Regression, SVM, KNN, Random Forest, Deep Learning, CNN.*

I. INTRODUCTION

CANCER has been one of the root causes of death worldwide. According to WHO (World Health Organization) 9.6 million deaths in 2018 were due to cancer [1]. Cancer is the term used when the cells in our body grow abnormally beyond their boundaries and slowly invade other parts of our body. The most common cases of Cancer as maintained by the IARC (International Agency for Research on Cancer) are Lung, Breast, Colorectal, Prostrate, Skin and Stomach. Breast Cancer reported 2.09 million cases in the year 2018 and caused 627000 deaths globally [2]. The SEER (Surveillance, Epidemiology and End Results) estimates 268,000 cases and 41,760 deaths due to Breast cancer in the year 2019[3].

The risk factors behind Breast cancer are unhealthy diet, lack of physical activity, excessive alcohol and tobacco intake, and leading a sedentary lifestyle. Fig.1 shows the incidences and mortality rate due to breast cancer globally. The cause of Breast cancer is still unclear hence prevention is the only possibility available. Thus, detection of the tumors in the breast till now is the sole way of curing this deadly disease.

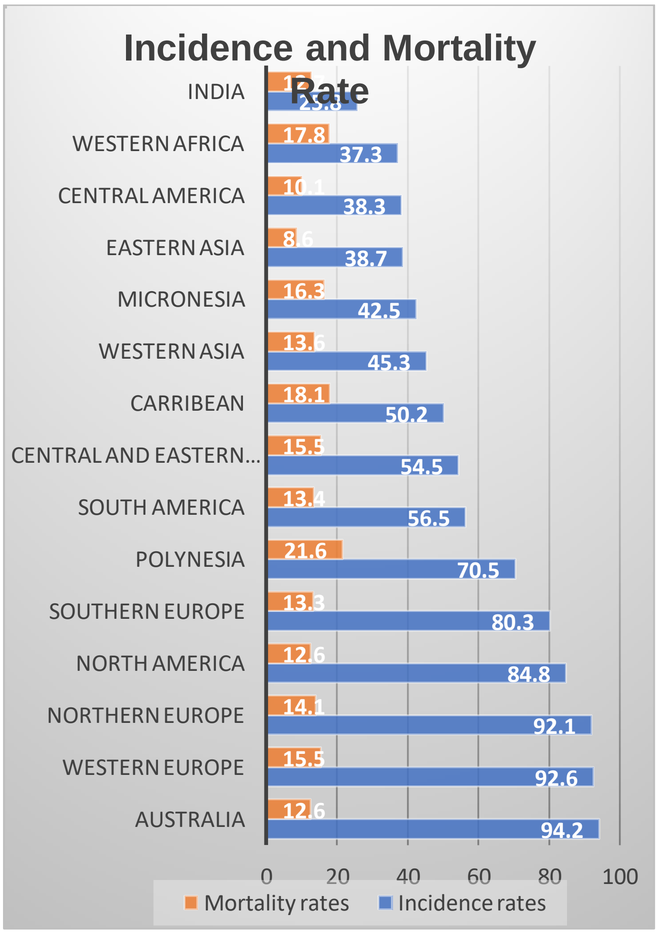


Fig. 1: The incidence and mortality rates worldwide [3]

The number of datasets available publicly is multitudinous. Following are some Breast Cancer datasets available in the internet:

1. Breast Cancer Wisconsin - https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
2. Breast Histopathology Images IDC - https://www.kaggle.com/paultimothymooney/predi

Revised Manuscript Received on May 06, 2019

Prof.Puspanjali Mohapatra, Department of computer science and Engineering, International Institute of Information Technology (IIIT), Bhubaneswar (Odisha), India.

Baldev Panda, Department of Electronics & Telecommunication, International Institute of Information Technology (IIIT), Bhubaneswar (Odisha), India.

Samikshya Swain, Department of Electronics & Telecommunication, International Institute of Information Technology (IIIT), Bhubaneswar (Odisha), India.

ct-idc-in-breast-cancer-histology-images/data.

3. Breast Cancer Histopathological database(BreakHis)-https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
4. BreCaHAD- A dataset for Breast cancer Histopathological annotation and diagnosis -https://figshare.com/articles/BreCaHAD_A_Dataset_for_Breast_Cancer_Histopathological_Annotation_and_Diagnosis/7379186 .

Advances in the Medical Science have led to adaptation of latest technologies in Breast cancer diagnosis and its treatment. A number of diagnosing methods are commonly used like CT scan, PET scan, MRI, Ultrasound, Mammograms and Histopathological Analysis [4]. Histopathology Analysis uses both software and hardware to extract the important features of the tissue and then prepares them for the image analysis. The tissues are mounted on slides and examined under microscopes to detect any growth in cancer cells or genetic progression. With the advent of WSI (Whole-Slide Imaging) scanners, the tissue Histopathological slides can be digitized and generated into a computerized image. The main motive of these digitized techniques is to obtain the quantitative data like cell size, tissue abnormalities and uneven number of cells.

The steps involved in the Image Analysis using Histopathology include:

1. Pre-Processing: This is the stage where operations like low pass filtering, dilation, thresholding are carried out [5].
2. Segmentation: Significant regions are extracted along with the ROI of the image to separate the image from the background. Some common techniques include: HMM, ACM, Watershed Algorithm [6].
3. Feature Extraction and Classification: The visual information of the image is obtained with the help of Feature Extraction. These are the input to be fed for classification. Classifiers like ANN, CNN learn during the training phase and then classify the cancerous nuclei into diverse classes during the testing.

   These methods can be time-consuming and restrictions like human error, bad image quality and misdiagnosis can cost someone's life. CAD (Computer Aided Diagnosis) is one recent context in radiology. CAD helps in improving the performance of the pathologists and the radiologists in finding out the cancerous tissue. A needle might not be able to extract as much information as a computer-aided system can to detect an abnormal tissue with improved accuracy and precision. Provided with the efficacious image processing steps a high level of efficiency can be achieved in cancer diagnosis with the help of automated diagnosis. Fig. 2 illustrates the overall process in CAD.

As we deal with a huge dataset of patients worldwide, applying conventional machine learning algorithms would not be a good recommendation. This is where Deep Learning plays a key role. Deep learning handles huge dataset perfectly and can also extract high-level features without any domain interference or hard-core feature extraction. Deep learning takes a long time to train but the testing phase is faster than a machine learning algorithm which is the reason it is preferred for handling complex problems like image classification, speech recognition, NLP easily.

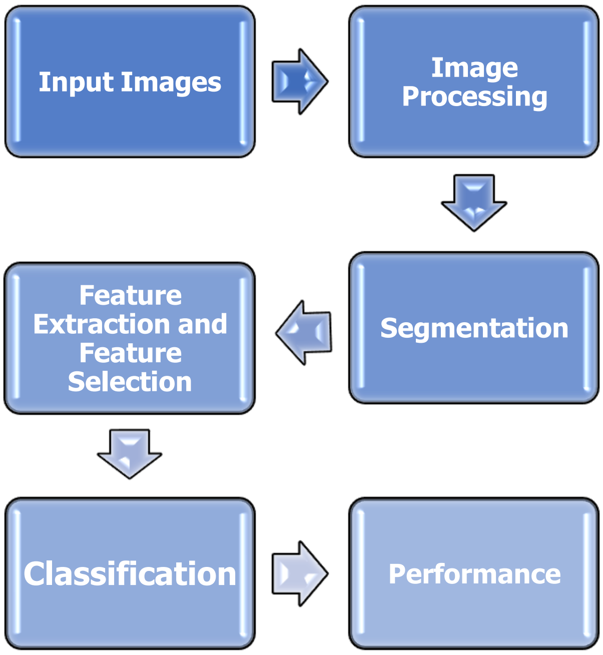


Fig. 2: Overall process of CAD in breast cancer diagnosis

The remaining paper has been structured as follows: Section II describes the various Datasets used in the analysis. Section III describes the related work in the field. Section IV gives an introduction to the Deep learning approach using CNN, Section V describes the proposed CNN Deep Learning Neural network in Image Classification. Section VI reports the experimental results and discusses the result analysis. Finally, the work is concluded, giving some insight for future work.

II. Dataset Description

The work in this paper uses 3 datasets for the cancer detection in Breast Histopathological Images.

The first analysis is carried out on a dataset which is basically small patches of images extracted from the Invasive Ductal Carcinoma (IDC) Breast Histopathological Image dataset. It can be downloaded from https://www.kaggle.com/simjeg/lymphoma-subtype-classification-fl-vs-cll . This dataset contains 5547 histology images of size 50x50x3 (width, height, channels) in the form of two numpy arrays X.npy and Y.npy. The patches have been labeled as 0 for No IDC and 1 for presence of IDC. The total number of Non–IDC images is 2759 and number of IDC detected images is 2788. The work described in this paper first applies the models on this dataset and the model which gives the maximum accuracy is selected further for classification on the original IDC dataset consisting of 277,524 images.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The dataset that applies the model selected from the previous dataset is the Invasive Ductal Carcinoma (IDC) Breast Histopathological Image which can be downloaded from-https://www.kaggle.com/paultimothymooney/predict-idc-in-breast-cancer-histology-images/data.

The dataset was originally organized by Janowczyk and Madabhushi and Angel Cruz- Roa [7] but it can be publicly accessed from the provided link. The original dataset consists of 162 whole mount slide images of H&E stained breast histopathology samples. These samples were scanned at 40x. Slide images due to their large spatial dimension are difficult to work with hence from these samples 277,524 patches of size 50 x50 were extracted out of which:

198,738- Negative Samples (No cancer)

78,786- Positive Samples (Cancer found in patches)

These patches that include IDC are labeled 1 and patches that don't include IDC are labeled 0. The dataset clearly shows that there are 2x negative samples than the positive samples.

Each image in the dataset has a specific file format:

u_xX_yY_classC.png->10253_idx5_x1351_y1101_class0.png

u – Patient's ID

X – X coordinate of the patch from where it was cropped

Y- Y coordinate of the patch from where it was cropped

C – Represents the class 0 or 1 (IDC or non-IDC)

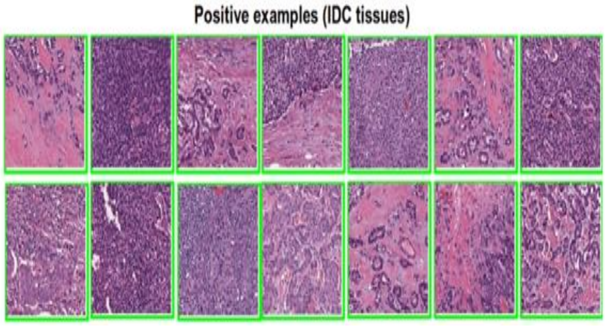


Fig. 3: Positive samples of the breast histopathology dataset

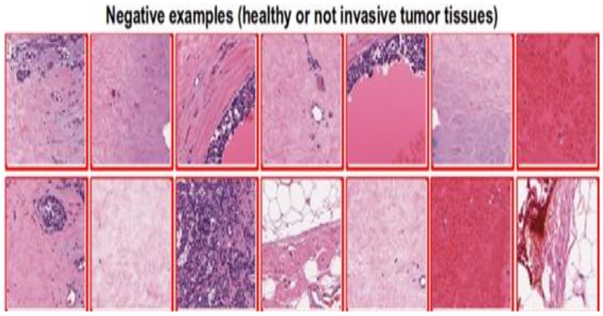


Fig. 4: Negative samples of the breast histopathology dataset

In this paper 120,000 samples of images have been used to train the model and the results have been plotted accordingly. Finally, based on the accuracy and precision obtained from the former datasets Deep learning is applied on a dataset containing high-resolution histopathology images. The Breast Cancer Histopathological Image Classification (BreakHis) dataset https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ is composed of 7909 images of breast cancer tissues. The data was collected from 82 patients with varying magnification factors (40x,100x,200x,400x). The dataset contains 2,480 benign and 5,429 malignant patches of images in PNG format. They are of resolution 700x460 pixels ,3-channel RGB with 8-bit depth. The dataset was curated in collaboration with the P&D lab –Pathological, Anatomy and Cytopathology, Brazil [8]. This dataset contains four different classes of benign tumors – Tubular Adenoma (TA), Fibroadenoma(F), Adenosis(A) and Phyllodes Tumor(PT) as well as four classes of malignant tumors – Carcinoma(DC),Lobular Carcinoma(LC), Mucinous Carcinoma (MC) and Papillary carcinoma(PC).

Each image in the dataset has a specific file format:

SOB_B_TA-14-4659-40-001.png

and follows the notation:

<BIOPSY_PROCEDURE>_<TUMOR_CLASS>_<TUMOR_TYPE>-<YEAR>-<SLIDE_ID>-<MAG>-<SEQ>

MAGNIFICATION	BENIGN	MALIGNANT	TOTAL
40X	652	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
TOTAL IMAGES	2480	5429	7909

Table 1: Structure of BreakHis Dataset

III. RELATED WORK

The image classification for cancer diagnosis has been a research topic for more than 40 years, but it's still a challenging task due to the complexity of images. Filipczuk et al. [9] performed the classification task on cytological images of fine needle biopsies, to distinguish the image into benign or malignant, with 25-dimensional feature vectors & using four different types of classification and achieved an accuracy of 98% on 737 samples. Similar to [9], George et al. [10] proposed diagnosis for BS based on nuclei segmentation of cytological images, using different neural nets and support vector machines, achieved accuracy rate ranging from 76% to 94% on a dataset of 92 images. Fabio Alexander Spanhol applied Convolution Neural Network (CNN) [11] classifier on BreakHis dataset which consists of 9109 samples of different images of different magnification factors. The dataset was divided into training (70%) and testing (30%). CNN was applied and the accuracy was found to be about 88% [4]. This CNN model was trained on NVIDIA^{R}Tesla^{R} K40m GPU.

Most of the recent works on breast cancer classifier are focused on WSI (Whole-Slide-Imaging) [12],[13].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

However, the adoption of WSI & other forms of digital pathology are still facing many obstacles such as cost constrains, insufficient productivity for high-volume of clinical routines, intrinsic technology corners etc.

IV. THE DEEP LEARNING APPROACH

Deep Learning is a subfield of Machine Learning which uses supervised, unsupervised or semi-supervised learning to automatically form useful information from data. It is similar in structure and function to the human nervous system which processes complex information with the help of a compounded network of interconnected computational units. The Deep learning Network outshines the traditional classification algorithms. The conventional algorithms require the domain knowledge in creating certain feature extractors to minimize the complex data and make the patterns easy for the learning algorithm to work. The performance of the conventional machine learning algorithms stays on the rear side when compared with Deep Neural Networks while handling a very large dataset. Deep Neural Net requires high end mechanism like GPUs to deal with large matrices and their multiplication, contrary to the conventional techniques which don’t deal with such high-end matrix multiplications. CNN (Convolutional Neural Network) is a type of deep learning that works on feed-forward propagation. CNN finds its applications in various areas like image recognition, face recognition, speech recognition, image processing, object recognition etc.

The network of nerve cells in our brain transfer the electrical impulses from one neuron to another. This complex network of neurons where the Dendrites carry the impulses from the synapse to the nucleus of the cell, Soma where it is processed and then transferred to the Axon. The Axon then transfers the impulses to the Synapse which then pass it on to the Dendrites of the second neuron. Fig.5. shows the neuron in the human brain.

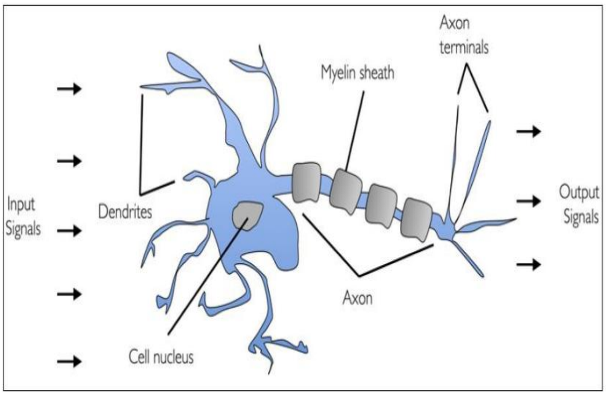


Fig. 5: The Biological Neuron of Human Brain

The computational concept in Deep Learning is inspired by the human brain. Multiple training layers form the Artificial Neural Network (ANN) which comprises layers of processing elements highly interconnected to manage the information.

Artificial neural networks have found extensive utilization due to their information processing capability pertinent mainly to non-linearity, parallelism and generalization techniques [14].

Some of the commonly used neural networks are listed below

NEURAL NETWORKS	POTENTIAL APPLICATIONS
ANN	Computational Neuroscience
CNN	Image Processing
RNN	Speech Recognition
DNN	Acoustic Modeling
DBN	Drug Discovery

Table 2: Commonly used neural networks

The neural networks have bias and weights which are updated according to the error. The weights are randomly initialized and are updated as the training proceeds. The bias is added after the weight multiplication which basically changes the range of the input weighted. The final linearly transformed input is generated which is then fed to the activation function. The activation function converts the input to output signals.

The commonly used Activation functions are:

1. Sigmoid- It generates a range of values between 0 and 1. It is defined as:

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad \ldots \text{(i)}$$

2. ReLU – It helps to train faster because it generates a constant derivative for all values greater than 0. The output is X if X>0 and 0 otherwise. It is defined as:

$$f(x) = \max(x, 0) \quad \ldots \text{(ii)}$$

3. SoftMax – This function is used when we have multiclass problems where the outputs are normalized so that the sum is 1.

A. Convolutional Neural Network

It is the neural network which is used in image recognition, object recognition, image classifications, face recognition etc. In CNN the input image is passed through a series of convolutional layers with kernels, pooling layer and fully connected layer before the final image classification. Convolution layer is the first stage where the features are extracted from the input image. Convolution is performed with different filters to perform functions like edge detection, sharpening or blurring.

When the filters cannot accommodate the input image, padding is used so that the output image is same size as the input image. There are two types of padding commonly used 1) Zero padding and 2) Valid Padding. The input image is then passed through a non-linear activation function like ReLU or Sigmoid or Tanh Function. When the number of parameters in a large image is high, pooling is used to reduce the dimension retaining the important information.

Pooling can be 1) Max pooling 2) Sum pooling 3) Average pooling. The last step includes feeding out image into a Fully Connected layer after flattening it into a vector. Finally, the flattened image is passed through the activation function to classify the output.

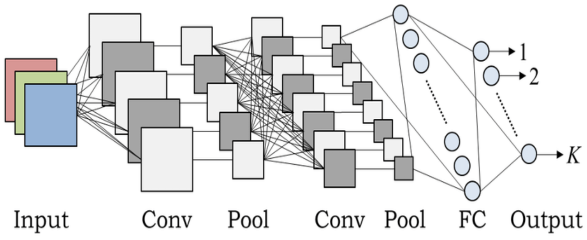


Fig. 6: CNN Architecture

CNN has found its application in Computer vision for Computer-Aided Detection [15]. Its most common application in our everyday life is facial recognition in Facebook where CNN detects the face of your acquaintances in a picture [16].Advances like SPPnet [17] and fast R-CNN [18] have drastically reduced the running time of object detection computations [19].From ALexNet [20] to ZF Net [21] , and then VGG Net [22], the ResNet [23] , the architecture of CNN is constantly improving. Not only these, CNN has found its applications in NLP (Natural Language Processing) [24] and have been extensively used to solve NLP tasks such as name entity recognition [25] and semantic role labeling [26].

V. IMAGE CLASSIFICATION USING CNN

The challenging task for image classification particularly the microscopic images from histopathological section is due to the large amount of inter-interaction variables, presence of complex geometrical structure, complex textures and minute details in image which can be the region of interest for classification [27]. Figure 7 depicts the complex texture found in histopathological images. Here Deep learning provides the possibilities of learning features directly from input data and process it through its hidden layers.

In this paper, the CNN model is used to classify image samples into cancerous or non-cancerous tissues and the accuracy is compared with other classifier models

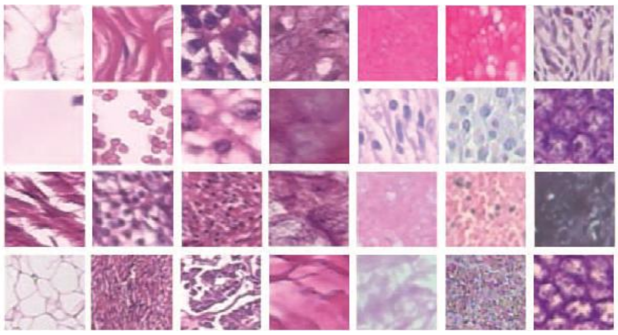


Fig. 7: Samples of real textures found in Histopathological images (H&E staining)

In deep learning, CNN plays a crucial role for processing visual-related problems [28].

Firstly, with a small Breast Histology Image dataset, consisting of 5547 samples of RGB digital images is given as input to the network with different architectures to find the best performing model. This step is called model selection step. Then the selected model is applied to the Invasive Ductal Carcinoma (IDC) Breast Histopathological Image dataset to observe the accuracy of the model.

The data is scaled from 0 to 1, so as to make the data compatible with wide variety of different classification algorithms. Fig. 8 shows the histogram plot of the IDC dataset.Also 20% of the data was set aside for K-fold cross-validation testing in Breast Histopathology Image dataset (NumPy array). This will make the model lesser prone to overfitting. As discussed earlier, there are four main layers used to build CNN architecture: Convolution layer, ReLU Layer, pooling layer, fully-connected layer. Normally, a full CNN architecture is obtained by stacking several of these layers. Different machine learning algorithms are applied on the Breast Histopathology Image dataset and the observations are recorded in Table 3. The maximum accuracy is found to be 73% for Random Forest and worst accuracy of 67% for Decision Tree classifier. The IDC dataset which contains 277,524 image samples are re-sized to 40X40 pixels from 50x50 and fed to the input of CNN to observe the accuracy.

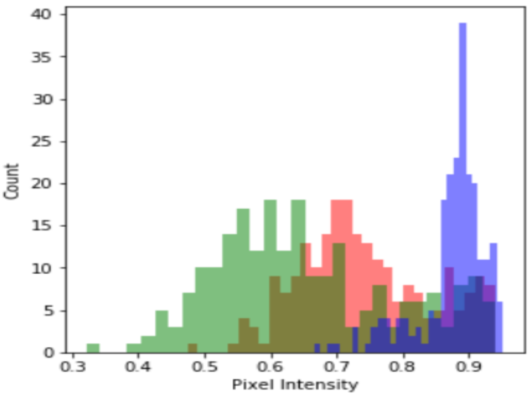


Fig. 8: Histogram plot of IDC Breast Histopathology data samples scaled between 0 and 1

Different architectures of CNN were applied to the dataset consisting of 5547 patches of images that were extracted from the original IDC dataset of 277,524 samples. The accuracies of different architectures are compared in Table 4. It can be observed that most of the architectures suffer from bias . The best performing model gives an accuracy of 76% with negilgible bias. The BreakHis dataset is split into two parts, training and validation sets. Different transform techniques like random scaling, cropping and flipping were applied on traing set. Since a pre-trained network RESNET-152 is used [29] , so the input image was resized as required by the network. Also normalization methods were applied on each color channel to centre it at 0 to 1 range.Based on the pre-trained network, new untrained feed-forward network acts as a classifier using ReLU as the activation function. The Adam optimizer with a learning rate of 0.005 is used for the training set. The total time taken to complete the training was 18 min 20s and the accuracy was found to be 89%.

The results have been plotted in Section VI.

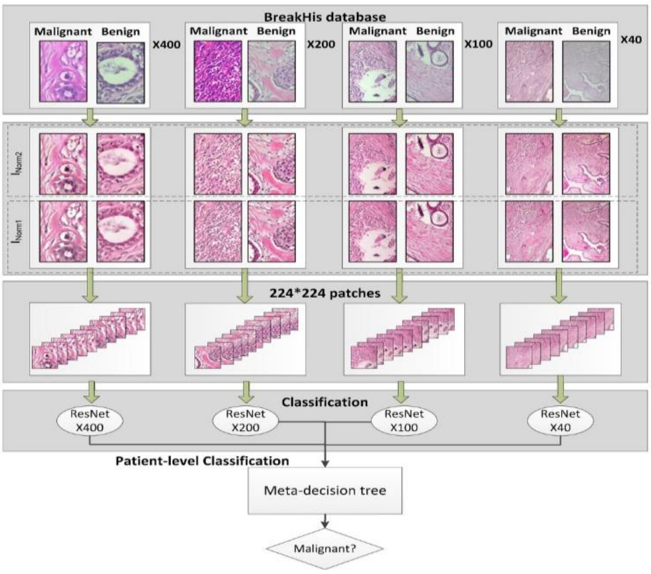
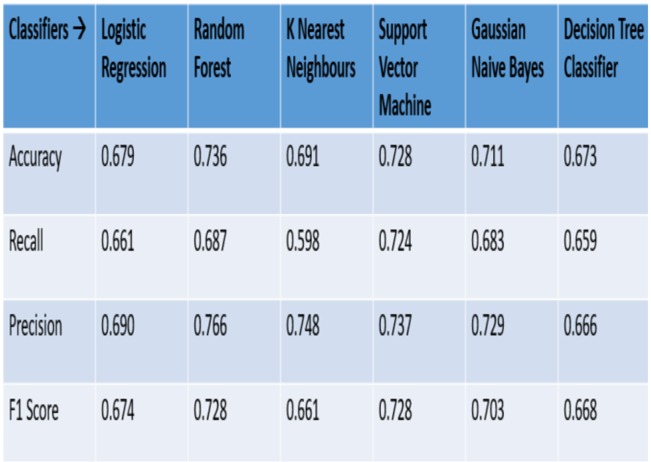


Fig.9. Framework of BreakHis Dataset



Classifiers →	Logistic Regression	Random Forest	K Nearest Neighbours	Support Vector Machine	Gaussian Naive Bayes	Decision Tree Classifier
Accuracy	0.679	0.736	0.691	0.728	0.711	0.673
Recall	0.661	0.687	0.598	0.724	0.683	0.659
Precision	0.690	0.766	0.748	0.737	0.729	0.666
F1 Score	0.674	0.728	0.661	0.728	0.703	0.668

Table 3: Comparisonof the performace of different classification algorithms applied on Breast Histopathology dataset (NumPy array).

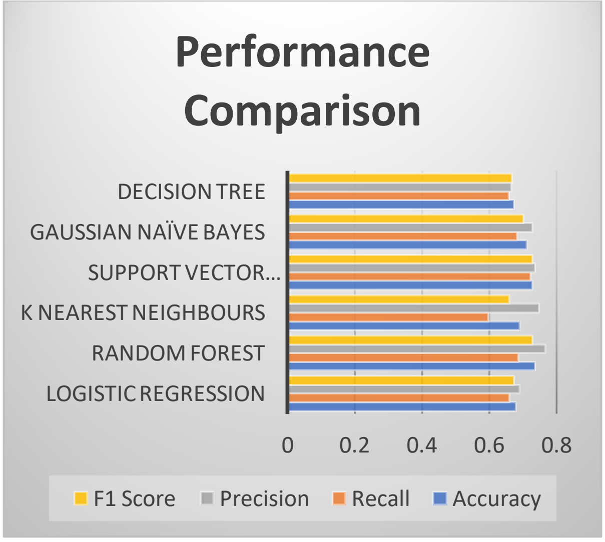


Fig.10. Plots for Comparison of different classifiers.

CNN ARCHITECTURE	ACC.	LEARNING CURVE
Conv(32)→Conv(64)→MaxPooling→Dropout→Flatten→Dense→Output	76%	Model has some bias
Conv(32)*2→MaxPooling→Conv(64)*2→MaxPooling→Conv(86)*2→MaxPooling→Dropout→Flatten→Dense→Dropout→Output	70%	Model Suffers from high Bias
Conv(32)*2→MaxPooling→Conv(64)*2→Maxpooling→Conv(86)→Max Pooling→Flatten→Dense→DropOut→Output (with Data Augmentation)	61%	Model Suffers from high Bias
Conv(32)→Conv(64)→MaxPooling→Dropout→Flatten→Dense→Drop out→Dense→Data Augmentation	76%	Best Performing model. Suffers no bias

Table 4: Classification accuracy comparison of various CNNarchitecture

The model is trained on CPU for 120000 (IDC Dataset) image samples out of which 96000 are used as training data sets and 24000 are test data sets.

A. Proposed Framework for CNN Architecture

In the end the CNN architecture that provides the best results for the experiment contains following layers & parameters:

- Input Layer: The input images are loaded to this layer & the produced output is used to feed the convolution layers. Several image processing techniques such as resize of the image to 40x40 pixels from 50X50 are applied. In our case the dataset consists of the image and parameters are defining the image dimensions (50X50 pixels) and the number of channels is 3 for RGB.
- Convolutional Layer: In this layer the convolution between input images with a set of learnable features or filters are carried out. By finding rough feature matches, in the same position for two images, CNN gets a lot better at seeing the similarity than whole image matching schemes.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

There are two convolution layers in our model with the receptive fields or kernel size of 3X3, the stride is set to 2. The first convolution layer learns 32 filters and second one learns 64 filters and it is initialized from a gaussian distribution with standard deviation of 0.0001.

- ReLU Layer: This layer removes the negative values from the filtered images and replaces it with zeros. For a given input value y, the ReLU layer computes output f(y) as y if y>0, else 0 or simply the activation is threshold at zero.
- Pooling Layer: In this layer the image stack is shrinked into a smaller size. The chosen window size is 2X2 with a stride of 2. For moving each window across the image, the maximum value is taken.
- Fully Connected Layer: This is the final layer where the actual classification happens. All filtered &shrinked images are stacked up. It passes the flattened output to the output layer where SoftMax classifier is used to predict the labels.

The dropout layers are added after the convolution & pooling layers to overcome the problem of overfitting to some extent. In our case it is 0.25 and 0.5. This layer arbitrarily turns off a fraction of neurons during the training process, which reduces the habituation on the training set by some amount. The fraction of neurons to turn off is decided by a hyperparameter, which can be tuned accordingly. The loss function used here is categorical cross-entropy and the keras optimizer used is Adadelta. Data Augmentation methods are also applied for image data such as horizontal & vertical flip, range of rotation, width and height shift etc. It is required to obtain more data for training in general and add it to the training set, also used to reduce overfitting.

VI. RESULTS AND DISCUSSION

The best performing architecture of CNN is selected according to the classification accuracy and the images are given as input for training with 8 epochs and batch size of 128.

The learning curve of the IDC Dataset for training and validation set is plotted in Fig. 11 and Fig.12. It can be observed that the model is not overfitted. The learning curve shows that the training and validation losses are gradually decreasing and training and validation accuracies are gradually increasing upon each epoch. The classification report of the model is shown in table 5. This model is able to achieve an accuracy of about 81% for 120000 image samples. Comparing the classification accuracies, the radar plot (Fig.13) shows that the training accuracy and validation accuracy are nearly equal over 10 epochs.

Also in Fig. 14 the training and validation losses are found to be nearly equal over 10 epochs.

	IDC(-)	IDC(+)
Precision	**0.72**	**0.89**
Recall	**0.92**	**0.63**
F1-score	**0.81**	**0.74**

Table 5: Classification report of IDC dataset

A. Accuracy plot for IDC dataset (trained on CPU)

Epoch	Training Accuracy	Validation Accuracy
1	0.73	0.78
2	0.78	0.79
3	0.79	0.80
4	0.80	0.79
5	0.80	0.81
6	0.81	0.81
7	0.80	0.80
8	0.81	0.82

Table 6: Table showing Training and Validation accuracy for IDC dataset

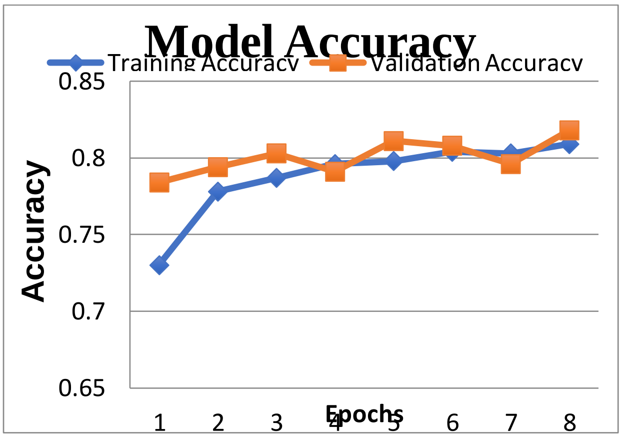


Fig.11. Accuracy Plots for training and Validation Sets

B. Loss plot for IDC dataset (trained on CPU)

Epoch	Training Loss	Validation Loss
1	0.55	0.51
2	0.49	0.45
3	0.47	0.44
4	0.46	0.46
5	0.45	0.42
6	0.44	0.42

7	0.43	0.44
8	0.42	0.41

Table 7: Table showing Training and Validation loss for IDC dataset

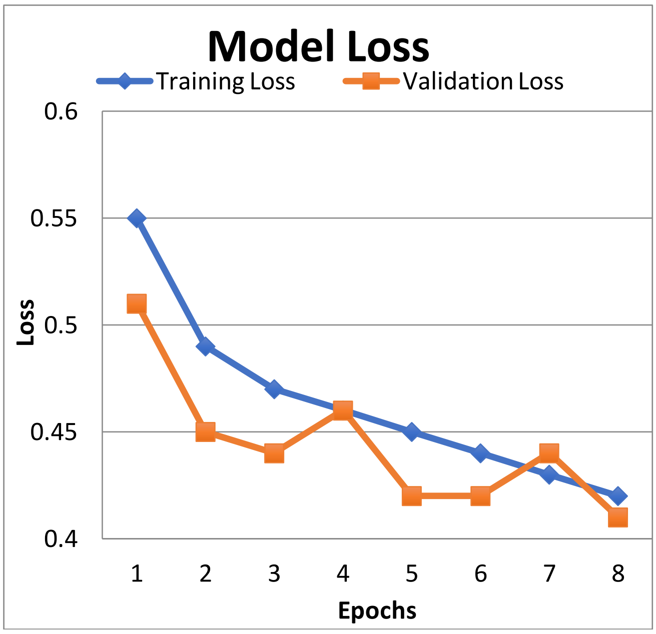


Fig.12.Model Loss performance for training and Validation sets

C. Accuracy plot for BreakHis dataset (trained on GPU)

Epoch	Training Accuracy	Validation Accuracy
1	0.749	0.760
2	0.832	0.835
3	0.855	0.866
4	0.870	0.858
5	0.875	0.889
6	0.882	0.885
7	0.883	0.884
8	0.889	0.884
9	0.882	0.886
10	0.881	0.891

Table 8: Table showing Training and Validation accuracy for BreakHis dataset

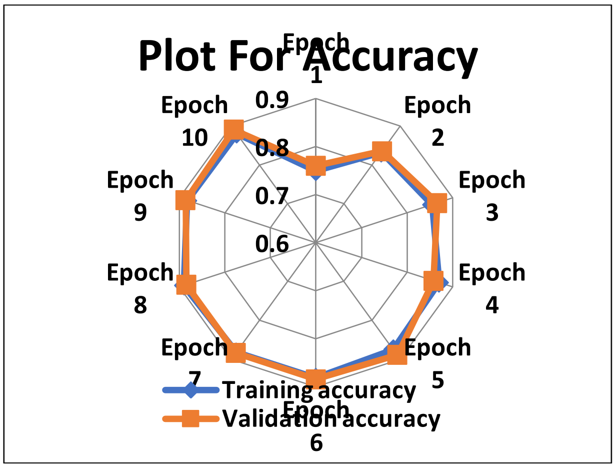


Fig.13. Accuracy Plots for training and Validation sets

D. Loss plot for BreakHis dataset (trained on GPU)

Epoch	Training Loss	Validation Loss
1	0.527	0.486
2	0.418	0.401
3	0.361	0.348
4	0.334	0.344
5	0.318	0.309
6	0.303	0.3
7	0.296	0.305
8	0.298	0.31
9	0.301	0.3
10	0.298	0.294

Table 9: Table showing Training and Validation loss for BreakHis dataset

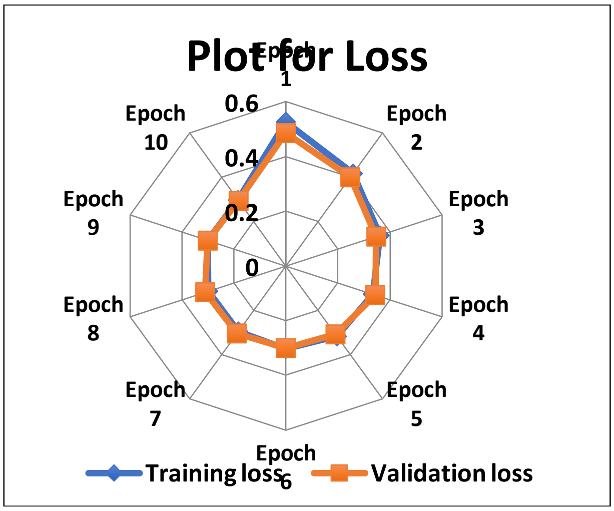


Fig.14. Loss Plots for training and Validation sets

Initially the models are trained in CPU.GPU (General Processing Unit) has been used for high-resolution larger dataset. GPU has large number of simple cores, which allows parallel computing through thousands of threads computing at a time [30].

VII. CONCLUSION

In this paper, Machine Learning and Deep Learning approaches have been implemented for Breast cancer classification. Different ML algorithms like Logistic Regression, KNN, SVM and Decision Tree are discussed and their accuracies are compared. Deep learning approaches like CNN are also studied and various performance measures have been evaluated to study the accuracy of the best architecture. The best performing CNN architecture gives an accuracy of 81% which is far more superior than the conventional ML algorithms. Larger dataset consisting of multiple resolutions of images were trained on GPU leading to an improved performance accuracy of 89%. This paper indicates that Deep learning approaches can efficiently classify the breast cancer samples compared to other models discussed in the paper.As a scope of future work high-resolution images can be trained using GPU like CUDA toolkit or Google Colab which supports free GPU. Also, different CNN architectures can be explored and optimization of the hyper parameters can be done.

REFERENCES

1. J. Ferlay, I.Soerjomataram , M. Ervik , R. Dikshit , S. Eser ,C. Mathers et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11Lyon, France: International Agency for Research on Cancer; 2013
2. F. Bray, J. Ferlay, I. Soerjomataram, R.L.Siegel,L.A. Torre, and A. Jemal, 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), pp.394-424.
3. Siegel, L. Rebecca, Kimberly D. Miller, and J. Ahmedin. "Cancer statistics, 2019." CA: a cancer journal for clinicians69.1 (2019): 7-34.
4. Bonnema, Jorien, et al. "Ultrasound-guided aspiration biopsy for detection of nonpalpable axillary node metastases in breast cancer patients: new diagnostic method." World journal of surgery 21.3 (1997): 270-274.
5. Ponraj, D. Narain, et al. "A survey on the preprocessing techniques of mammogram for the detection of breast cancer." Journal of Emerging Trends in Computing and Information Sciences 2.12 (2011): 656-664.
6. M. A., Aswathy,and M. Jagannath. "Detection of breast cancer on digital histopathology images: Present status and future possibilities." Informatics in Medicine Unlocked 8 (2017): 74-79.
7. Cruz-Roa, Angel, et al. "High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection." PloS one 13.5 (2018): e0196828.
8. F.Spanhol, L. S.Oliveira, C. Petitjean, L. Heutte, A Dataset for Breast Cancer Histopathological Image Classification, IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016.
9. Filipczuk, Paweł, et al. "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies." IEEE Transactions on Medical Imaging32.12 (2013): 2169-2178.
10. George, Yasmeen Mourice, et al. "Remote computer-aided breast cancer detection and diagnosis system based on cytological images." IEEE Systems Journal 8.3 (2014): 949-964.
11. Spanhol, Fabio Alexandre, et al. "Breast cancer histopathological image classification using convolutional neural networks." 2016 international joint conference on neural networks (IJCNN). IEEE, 2016.
12. Zhang, Yungang, et al. "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles." Machine vision and applications 24.7 (2013): 1405-1420.
13. Farahani, Navid, Anil V. Parwani, and LironPantanowitz. "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives." Pathol Lab Med Int 7 (2015): 23-33.
14. Andrews, Robert, Joachim Diederich, and Alan B. Tickle. "Survey and critique of techniques for extracting rules from trained artificial neural networks." Knowledge-based systems8.6 (1995): 373-389.
15. Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285-1298.
16. Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
17. S. Zagoruyko, &N. Komodakis, (2015). Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4353-4361).
18. Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.
19. Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
20. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
21. Du, Xuedan, et al. "Overview of deep learning." 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE, 2016.
22. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
23. Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures." arXiv preprint arXiv:1603.08029
24. Young, Tom, et al. "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine 13.3 (2018): 55-75.
25. Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." LingvisticaeInvestigationes 30.1 (2007): 3-26.
26. Vickrey, David, and Daphne Koller. "Sentence simplification for semantic role labeling." Proceedings of ACL-08: HLT (2008): 344-352.
27. Mussurakis, Stavros, David L. Buckley, and Anthony Horsman. "Dynamic MRI of invasive breast cancer: assessment of three region-of-interest analysis methods." Journal of computer assisted tomography 21.3 (1997): 431-438
28. W. Hu, Y. Huang, L. Wei, F. Zhang, &H. Li, (2015). Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors, 2015.
29. Motlagh, NimaHabibzadeh, et al. "Breast cancer histopathological image classification: A deep learning approach." bioRxiv (2018): 242818.
30. Bergstra, James, et al. "Theano: Deep learning on gpus with python." NIPS 2011, BigLearning Workshop, Granada, Spain. Vol. 3. Microtome Publishing., 2011